DOCUMENT RESUME

ED 333 015 TM 016 463

AUTHOR	Ackerman, Terry A.; Davey, Tim C.
TITLE	Concurrent Adaptive Measurement of Multiple Abilities.
PUB DATE	Apr 91
NOTE	24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1991).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Ability; *Adaptive Testing; *Computer Assisted Testing; Equations (Mathematics); *Item Banks; *Mathematical Models; Test Validity
IDENTIFIERS	Ability Estimates; *Concurrent Adaptive Measurement; *Unidimensionality (Tests)

ABSTRACT

An adaptive test can usually match or exceed the measurement precision of conventional tests several times its length. This increased efficiency is not without costs, however, as the models underlying adaptive testing make strong assumptions about examinees and items. Most troublesome is the assumption that item pools are unidimensional. Truly unidimensional item pools are the exception rather than the rule, so procedures have been established for handling multidimensional pools. One option is to insure that every adaptive test measures the same composite of the multiple abilities represented in the item pool. However, this approach forfeits the multidimensional structure of the item pool. The alternative is to retain this structure by splitting the item pool into more unidimensional subsets and administering each separately. This approach, however, increases testing time. A third approach is proposed--concurrent adaptive measurement. In this approach collateral information--information that an item provides about a secondary ability--is used to update ability estimates obtained from adaptive tests administered in separate content areas. A study is reviewed, which evaluated the effectiveness of the concurrent adaptive measurement procedure using unidimensional estimates of two two-dimensional item pools of 200 items each. The results indicate that both bias and the standard error of the estimated ability decrease when collateral information is used. As the correlation between the latent skills increases, the standard error drops slightly. Six tables and four figures are included. (SLD)

Concurrent Adaptive Measurement of Multiple Abilities

Terry A. Ackerman

University of Illinois

Tim C. Davey

American College Testing Program

Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April, 1991

Abstract

An adaptive test can usually match or exceed the measurement precision of conventional tests several times its length. This increased efficiency is not without costs, however, as the models underlying adaptive testing make strong assumptions about examinees and items. Most troublesome is the assumption that item pools are unidimensional. Because truly unidimensional item pools are the exception rather than the rule, procedures have been established for handling multidimensional pools. One option is to insure that every adaptive test administered measures the same composite of the multiple abilities represented in the item pool. However, this approach forfeits information by losing the multidimensional structure of the item pool. The alternative is to retain this structure by splitting the item pool into more unidimensional subsets and administering each separately. One major drawback with this approach is the increase in testing time. In this paper we propose a third approach called ***concurrent adaptive measurement***. In the new approach ***collateral information***, information that an item provides about a secondary ability, is used to update ability estimates obtained from adaptive tests administered in separate content areas.

Concurrent Adaptive Measurement of Multiple Abilities

Efficiency is one of the principle advantages afforded by computerized adaptive testing (CAT). Conventional tests can waste time by presenting examinees items that are either much too easy or much too difficult. These inappropriate items contribute essentially no measurement information and thus can be excluded without affecting test quality. Exclusion of inappropriate items is the basic tenet of adaptive testing, where item and test difficulty are tailored to the examinee's level of ability. By presenting only appropriate items, a relatively short adaptive test is able to match the measurement precision of conventional tests many times its length.

The increased efficiency of adaptive testing is not without cost, however, as the models underlying CAT make strong assumptions about examinees, items, and their interaction. Most troublesome is the assumption that item pools are unidimensional, or that each item is measuring the same trait or exact composite of multiple traits. This assumption is rarely tenable, with unidimensional pools the exception rather than the rule. Most item pools can be broken down into more homogeneous subsets, each more nearly a unidimensional measure of a narrower, more specific trait. For example, items collectively designed to measure mathematics proficiency can be further classified as measures of numerical operations, algebra, geometry, and so on.

Item response theory (IRT) models have been found to be fairly robust to the kind of hierarchical multidimensionality typically observed in a broad test of mathematics (Reckase, 1979; Drasgow & Parsons, 1985). However, their applications to adaptive testing may still be compromised since unidimensionality is critical when different examinees are administered different sets of items. Ackerman (in press) demonstrated that if tests administered in a CAT are not constrained to sample similar proportions of each content area obtained scores may not be comparable across examinees. Using a multidimensional mathematics pool, Ackerman demonstrated that it is possible for examinees at one level of ability to be presented primarily items measuring numerical skills, while examinees at a higher level of ability might receive only geometry items because of their higher level of difficulty. In essence, these two groups of examinees were administered tests of different content, and thus, their scores are not directly comparable.

Several different approaches can be used to resolve the problems presented by multidimensional item pools. All begin by classifying pool items into more homogeneous categories, or content domains. Each of these domains is expected to provide a nearly unidimensional measure of a specific ability. One approach to test administration, called content balancing, insures that each adaptive test presented includes some fixed proportion of items form each content domain. For example, the complete test may be composed of 20% numerical skills, 50% algebra, and 30% geometry items. In this way, each adaptive test is constrained to measure the same unidimensional composite of the multidimensional item pool. Unfortunately, this approach forfeits information by losing the multidimensional structure of the item pool. The unidimensional composite fails to convey any variation across examinees in their profile of abilities in the more specific content domains.

An alternative is to treat domains separately and administer an adaptive test in each content. While this retains the multidimensional structure of the item pool, it does so at the cost of dramatically increasing administration time. Thus, the decrease in testing time usually afforded by CAT could be forfeited. Because content domains worth measuring are found to be nearly universally positively correlated, some small increase in efficiency may be afforded by using the ability estimates from previously administered domains to derive starting points in subsequent domains. However, evidence suggests that the gain in efficiency is small in most cases (Green & Thomas, 1990).

In this paper we outline a new approach, one that retains the multidimensional structure of the item pool without a serious loss of testing efficiency. The viability of the new approach, called ***concurrent adaptive measurement***, is demonstrated in a monte carlo simulation.

Concurrent adaptive measurement

Because content domains are correlated, an item from one domain will contribute measurement information about all other domains. This is most easily illustrated by taking, an item classified as measuring predominantly numerical skills and embedding it in a test containing only algebra items. The biserial correlation (or the IRT discrimination parameter, a) of the alien item will be reduced compared to its biserial value in it's proper context of the numerical skills domain, but it still will be nonzero. That is, each of the numerical skill items can provide information for estimating ability in the algebra domain and vice versa.

Administering each content domain independently wastes whatever information an item provides outside its own content domain. Through concurrent adaptive measurement, this information can be recovered by allowing every item presented to contribute to the estimation of ability in each domain. This requires that every item in the CAT pool have multiple sets of parameters--one for each content domain. These ***collateral*** parameters are computed by calibrating an item with respect to the trait defined solely by the members of a given content domain. One method of performing this calibration would be to first estimate IRT item parameters for each domain independently. Items outside the content domain are then included, one at a time, and calibrated while keeping the ***valid*** parameters of the "home" domain items fixed at their original estimates.

A more efficient method can be derived from the work of Wang (1986). Wang demonstrated analytically how two-dimensional item parameters and a two-dimensional underlying ability distributions get mapped into a unidimensional latent space. Essentially Wang's formulation enables one to compute 2PL IRT item parameters from a marginal item characteristic curve (ICC) that is computed in the direction of what she termed the reference composite.

The reference composite is based upon the first principle component of the $A\Omega A'$ matrix where A is a matrix of two-dimensional item discrimination parameters and Ω is the θ_1,θ_2 variance-covariance matrix. The angle associated with the reference composite defines the meaning of the unidimensional scale in terms of a θ_1-θ_2 composite. That is, if the reference composite angle was at 45^o the unidimensional θ scale could be interpreted as an equal weighting of θ_1 and θ_2 abilities.

Unidimensional parameters for a given item can be estimated from the computed marginal ICC in the direction of the reference composite. This ICC is a function of the item's two-dimensional parameters, the specified reference composite direction, and the characteristics (i.e., vector of θ_1 and θ_2 means, and the θ_1,θ_2 variance-covariance matrix) of the hypothesized two-dimensional underlying ability distribution. Valid unidimensional parameters would be obtained by using the reference composite for the item's home pool. Collateral item parameters are determined by using the using the reference composite of a pool of items measuring another correlated skill.

Once item parameters have been obtained, administration of the adaptive test can proceed in either of two ways. If the principle goal of testing is to produce a unidimensional composite estimate of proficiency for an entire pool, test administration would follow the content balancing model. However, in addition to the unidimensional composite, abilities could be further refined by using collateral parameters to estimate abilities within each content domain. This would yield a better estimate of the unidimensional composite, but somewhat less precise measures in the individual domains.

If, on the other hand, the principle goal of testing was an accurate estimation of each content domain, the domains could be presented sequentially and independently. The important difference between this approach and the pool splitting approach outlined above is that information would be accumulated in all content categories regardless of which is actually being presented. After the last content area has been administered final ability estimates for each content category would be the accumulation of the items administered in the specific category plus the information obtained via the collateral parameters from items administered in each of the other content areas.

Method

To evaluate the effectiveness of the concurrent adaptive measurement procedure the following study was conducted. Two two-dimensional item pools of 200 items each were created. The item parameters for Pool 1 were randomly selected with the constraint that the two-dimensional item vectors (c.f., Reckase, 1985) lie within a 30^{o} of the θ_1 axis. All of the item parameters for Pool 2 were the same as those in Pool 1 except the a_1 and a_2 two-dimensional discrimination parameters were reversed. Thus, the vectors for these item fell within 30^o of the θ_2 axis. The generated two-dimensional item parameters were reviewed and thought to be similar to those one would obtain in a typical two-dimensional IRT calibration of cognitive test data. A plot of the item vectors for each pool in the two-dimensional ability plane is shown in Figure 1.

Insert Figure 1 about here

The length of the item vector represents the amount of discrimination. The base of each vector is orthogonal to the item's p = .5 equiprobability contour and the angle the vector makes with the θ_1 axis represents the θ_1,θ_2 composite that is best measured by the item. The reference composite for each pool is illustrated by a dotted vector. The reference composite for Pool 1 had an angle of 15^o and for Pool 2, 75^o.

Valid and collateral unidimensional item parameters were obtained for each pool for each of three different θ_1,θ_2 correlational conditions, .3, .5, and .7. That is, for Pool 1 unidimensional item parameters were estimated from the marginal ICCs using the 15^o reference composite (i.e., the valid parameters) and using the 75^o reference composite direction (i.e., the collateral parameters). This process was repeated for each correlational level. The same process was used to obtain unidimensional estimates for the Pool 2 items.

For each pool the mean and standard deviation of the valid estimated item parameters were $\mu_{\hat{a}}=1.58$, $\sigma_{\hat{a}}=.49$, $\mu_{\hat{b}}=.00$, and $\sigma_{\hat{b}}=.79$. The mean and standard deviation of the estimated collateral item parameters were for r=.7: $\mu_{\hat{a}}=1.11$, $\sigma_{\hat{a}}=.29$, $\mu_{\hat{b}}=-.01$, and $\sigma_{\hat{b}}=.85$. r=.5: $\mu_{\hat{a}}=.92$, $\sigma_{\hat{a}}=.24$, $\mu_{\hat{b}}=-.05$, and $\sigma_{\hat{b}}=.86$. r=.3: $\mu_{\hat{a}}=.75$, $\sigma_{\hat{a}}=.21$, $\mu_{\hat{b}}=-.06$, and $\sigma_{\hat{b}}=1.00$.

Once the valid and collateral item parameters were obtained the two pools were used in a simulated CAT at each correlation level. In the CAT the most informative item was selected at the current estimated ability level. Bayes modal ability estimates (using a N(0,1) prior) were updated after each item was "administered". The number of items administered in each simulated CAT was fixed at 15.

Examinees were selected from a two-dimensional grid. One-hundred examinees were simulated at each of 81 θ_1,θ_2 levels (from -2.0 to +2.0 in increments of .5 along each ability dimension.) Thus, for each of the nine specified levels on the θ_1 ability scale there were 900 examinees simulated. The same was true for each specified level of θ_2.

The amount of information that was gained using the collateral item parameter estimates was assessed by examining the standard error of $\hat{\theta}$ and the bias $(\hat{\theta} - \theta)$ at each of the nine theta levels. That is, the difference in the bias and standard error values between abilities estimated with valid items only and abilities estimated with both valid items and colllateral items were compared.

Rather than using empirical estimates of these two measures it was decided to obtain theoretical estimators using the formulas suggested by Lord (1983). The standard error of $\hat{\theta}$ was calculated for each examinee using the IRT information function and the parameters of the items that were administered. At each θ level the information was averaged over all (900) examinees and then was used to compute the standard error via the formula

$$S.E.(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}.$$

Bias was computed using the formula (Lord, 1983, p.237)

$$Bias(\hat{\theta}) = \frac{1}{I^2}\sum_{i=1}^{n} a_i I_i (P_i - .5)$$

where a_i is the discrimination parameter of item i,

P_i is the probability of a correct response for the 2PL IRT model, and,

$I_i = \frac{(P_i')^2}{P_i Q_i}$ with P' being the first derivative of $P(\theta)$ with respect to θ.

The theoretical values were computed because it was believed that such an approach would provide an optimum comparison without introducing error accrued in the estimation of abilities.

Results

The Test 1 pool information function and the collateral information function on Test 1 from the Test 2 pool for each level of correlation are displayed in Figure 2. As might be expected, the amount of collateral information increases as the level of correlation increases. This occurs because as the correlation increases the two-dimensional latent space is, in a sense, collapsing into a single dimension, causing the reference composites to converge. Because the two-dimensional difficulty parameters were randomly selected from a N(0,1) distribution they produced a balance of easy and difficult items. Thus, the pool information function and well as the collateral information functions to be centered around $\theta = 0.0$.

Insert Figure 2 about here

The two estimates of bias for both Test 1 and Test 2 at each of the three correlational levels are displayed in Tables 1, 2 and 3. A graph of Table 3 is shown in Figure 3. Bias values based on the administration of 15 items from each pool are labeled as Test 1 and Test 2. Bias values that are based upon the estimated unidimensional parameters of both valid items and the collateral item parameters are labeled Test 1A and Test 2A.

It appears that the difference between the three correlational levels is quite negligible and perhaps due just to random fluctuation. However, the addition of collateral information did decrease the amount of bias at each θ level in every case. For every specified level of correlation bias was negatively correlated with θ. Zero bias occured at $\theta = 0$ for each correlational level.

Insert Tables 1, 2, & 3 and Figure 3 about here

The values of the theoretical standard errors for the two θ estimates at each of the three correlational levels for each are listed in Tables 4, 5, and 6. These tables are labeled in the same manner as their bias counterparts. As was expected the standard error of the condition θ distributions decrease as the θ_1-θ_2 correlations increased, although the decrease appeared to quite small. In a similar manner the ability estimates computed using the collateral information had a smaller standard error across all ability levels across all correlational conditions.

The inverse relationship between the IRT information function and standard error is clearly illustrated by comparing Figures 2 and Figure 4. Figure 4 is a plot of the standard error values listed in Table 6. The advantage of using collateral information is clearly illustrated.

Insert Tables 4, 5, and 6 and Figure 4 about here

Discussion

The purpose of this paper was to illustrate how collateral information from correlated tests can be used to improve CAT ability estimates. This study employed unidimensional estimates of two two-dimensional item pools, each measuring a different θ_1,θ_2 composite. Results indicate that both bias and the standard error of the estimated ability decrease when collateral information is used. Likewise, as the correlation between the latent skills increased, there was a drop in the standard error of $\hat{\theta}$, albeit quite small.

This study has presented an idea which takes advantage of the richness most items possess. It is suspected that most item pools are in actuality multidimensional and that representing them via unidimensional parameter estimates is limiting the amount of information that is available. In reality practitioners usually do not calibrate items to fit multidimensional models, but instead work with unidimensional estimates. More research needs to be done to help establish guidelines for using multidimensional analyses. It would be helpful if criteria could be established which would help the practitioner decide whether the increase in measurement precision through the use of collateral item information is warranted.

References

Ackerman, Terry (in Press) The use of unidimensional item parameter estimates of multidimensional items in adaptive testing. Applied Psychological Measurement

Drasgow, F. & Parsons, C.K. (1985). Application of unidimensional item response theory models to multidimensional data. In D.J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference 6. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1982.

Green, B.F., & Thomas, T.J. (1990). Utility of predicting starting abilities in sequential computerized adaptive tests. Presented at the Annual Meeting of the Psychometric Society, Princeton, NJ.

Lord, F.M. (1983) Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48, 233-245.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4 207-230.

Reckase, M.D. (1985, April). The difficulty of test items that measure more than one ability. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wang, M. M. (1986). Fitting a unidimensional model to multidimensional response data. A paper presented at the annual Office of Naval Research Contractor's Meeting. Knoxville, TN.

Table 1.
Bias values for abilities estimated using Test 1, Test 2, Test 1A, and Test 2A with θ_1,θ_2 abilities correlated .3.

Bias $(\hat{\theta} - \theta)$

θ	Test 1	Test 1A	Test 2	Test 2A
2.0	-.0777	-.0574	-.0776	-.0574
1.5	-.0521	-.0393	-.0526	-.0396
1.0	-.0232	-.0186	-.0230	-.0185
.5	-.0078	-.0066	-.0075	-.0064
.0	.0005	.0004	.0004	.0003
-.5	.0098	.0080	.0098	.0081
-1.0	.0224	.0179	.0222	.0177
-1.5	.0466	.0354	.0467	.0355
-2.0	.0855	.0622	.0859	.0625

Table 2.
Bias values for abilities estimated using Test 1, Test 2, Test 1A, and Test 2A with θ_1, θ_2 abilities correlated .5.

Bias $(\hat{\theta} - \theta)$

θ	Test 1	Test 1A	Test 2	Test 2A
2.0	-.0764	-.0575	-.0768	-.0577
1.5	-.0524	-.0388	-.0523	-.0388
1.0	-.0229	-.0180	-.0230	-.0180
.5	-.0076	-.0063	-.0076	-.0063
.0	.0001	.0001	.0012	.0008
-.5	.0097	.0078	.0102	.0080
-1.0	.0226	.0176	.0227	.0178
-1.5	.0467	.0352	.0461	.0348
-2.0	.0857	.0630	.0854	.0629

Table 3.
Bias values for abilities estimated using Test 1, Test 2, Test 1A, and Test 2A with θ_1, θ_2 abilities correlated .7.

Bias $(\hat{\theta} - \theta)$

θ	Test 1	Test 1A	Test 2	Test 2A
2.0	-.0774	-.0609	-.0767	-.0602
1.5	-.0522	-.0394	-.0524	-.0396
1.0	-.0229	-.0180	-.0228	-.0180
.5	-.0078	-.0064	-.0073	-.0062
.0	.0009	.0006	.0005	.0003
-.5	.0098	.0077	.0099	.0078
-1.0	.0225	.0178	.0222	.0177
-1.5	.0469	.0364	.0468	.0362
-2.0	.0860	.0659	.0858	.0659

Table 4.
Standard error values of abilities estimated using Test 1, Test 2, Test 1A, and Test 2A with θ_1, θ_2 abilities correlated .3.

$S.E.(\hat{\theta})$

θ	Test 1	Test 1A	Test 2	Test 2A
2.0	.4562	.2515	.4559	.2514
1.5	.3663	.2077	.3659	.2080
1.0	.2836	.1661	.2845	.1663
.5	.2381	.1412	.2383	.1413
.0	.2272	.1353	.2274	.1356
-.5	.2429	.1454	.2424	.1447
-1.0	.2945	.1755	.2941	.1756
-1.5	.3652	.2169	.3648	.2170
-2.0	.4568	.4571	.2704	.2711

Table 5.
Standard error values of abilities estimated using Test 1, Test 2, Test 1A, and Test 2A with θ_1, θ_2 abilities correlated .5.

$S.E.(\hat{\theta})$

θ	Test 1	Test 1A	Test 2	Test 2A
2.0	.4560	.2427	.4559	.2427
1.5	.3658	.2012	.3663	.2030
1.0	.2841	.1625	.2839	.1629
.5	.2386	.1391	.2383	.1397
.0	.2273	.1342	.2274	.1345
-.5	.2433	.1444	.2426	.1435
-1.0	.2941	.1742	.2943	.1744
-1.5	.3652	.2165	.3656	.2168
-2.0	.4567	.2703	.4569	.2700

Table 6.
Standard error values of abilities estimated using Test 1, Test 2, Test 1A, and Test 2A with θ_1, θ_2 abilities correlated .7.

$S.E.(\hat{\theta})$

θ	Test 1	Test 1A	Test 2	Test 2A
2.0	.4560	.2339	.4559	.2339
1.5	.3665	.1955	.3658	.1947
1.0	.2842	.1584	.2840	.1579
.5	.2385	.1369	.2384	.1372
.0	.2271	.1324	.2273	.1327
-.5	.2433	.1427	.2428	.1426
-1.0	.2943	.1739	.2946	.1736
-1.5	.3650	.2161	.3650	.2159
-2.0	.4569	.2714	.4571	.2711

Figure Captions

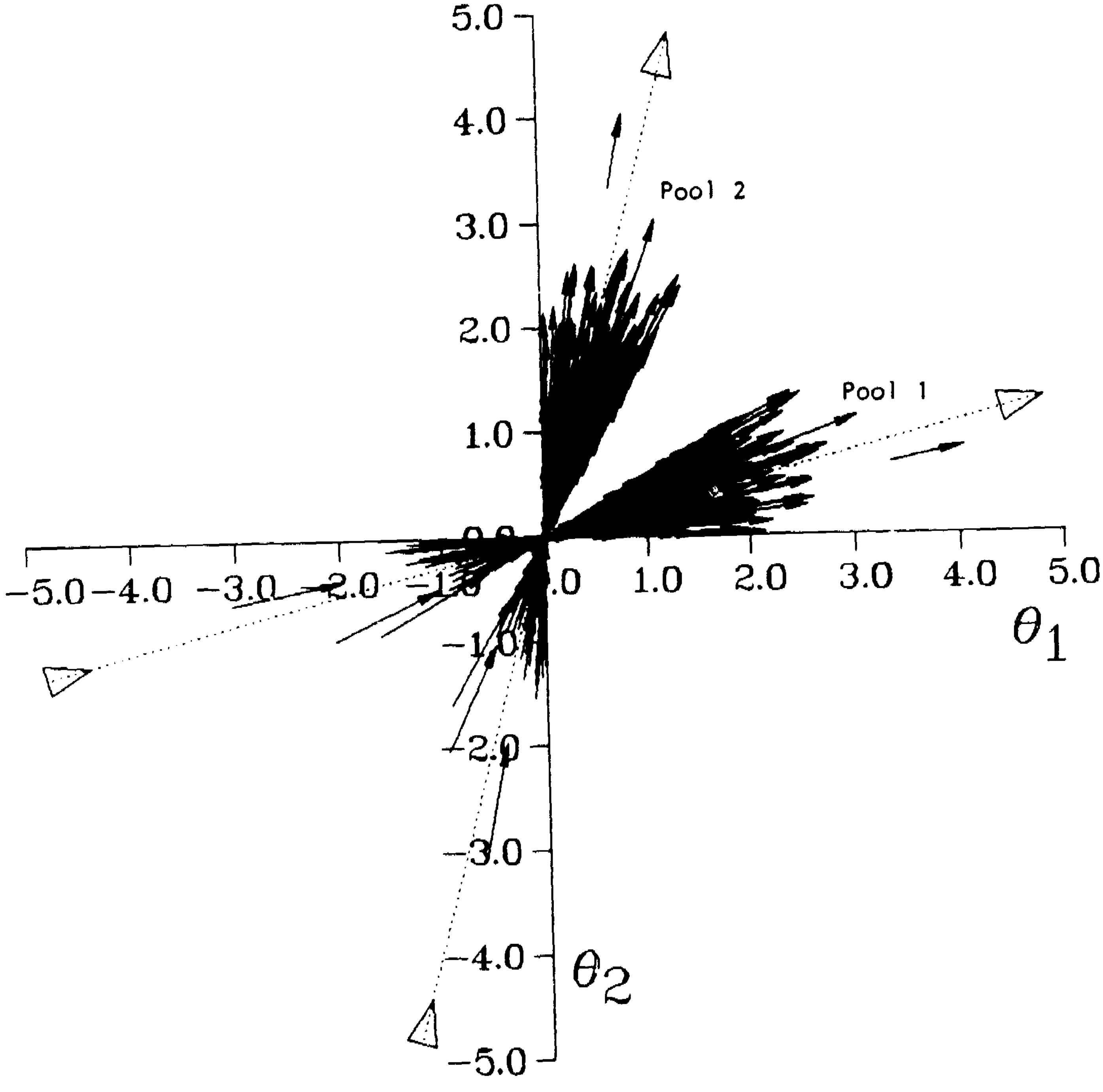
Figure 1. Two-dimensional item vectors representing items from Pool 1 and Pool 2.

Figure 2. The Pool 1 information curve and the three collateral information curves for r = .3, .5 and .7 from Pool 2.

Figure 3. Bias plot for Tests 1, 1A, 2 and 2a when the correlation between Test 1 and Test 2 is .7.

Figure 4. Empirical standard error plot for Tests 1, 1A, 2 and 2a when the correlation between Test 1 and Test 2 is .7.

Figure 1



5.0
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-5.0
-4.0
-3.0
-2.0
-1.0
1.0
2.0
3.0
4.0
5.0
Pool 2
Pool 1
θ_1
θ_2

Figure 2

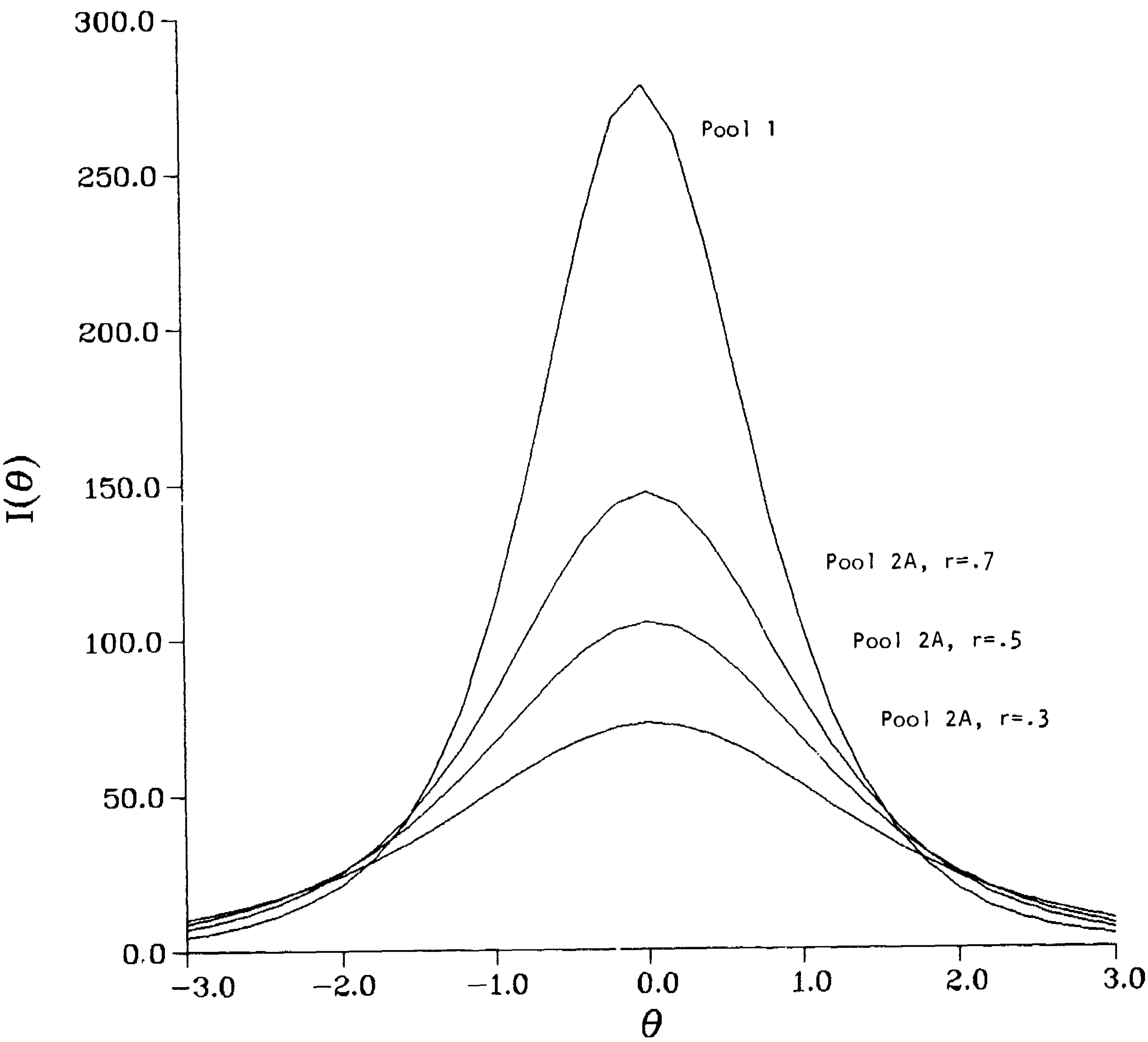

300.0
250.0
200.0
150.0
100.0
50.0
0.0
I(θ)
Pool 1
Pool 2A, r=.7
Pool 2A, r=.5
Pool 2A, r=.3
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 3

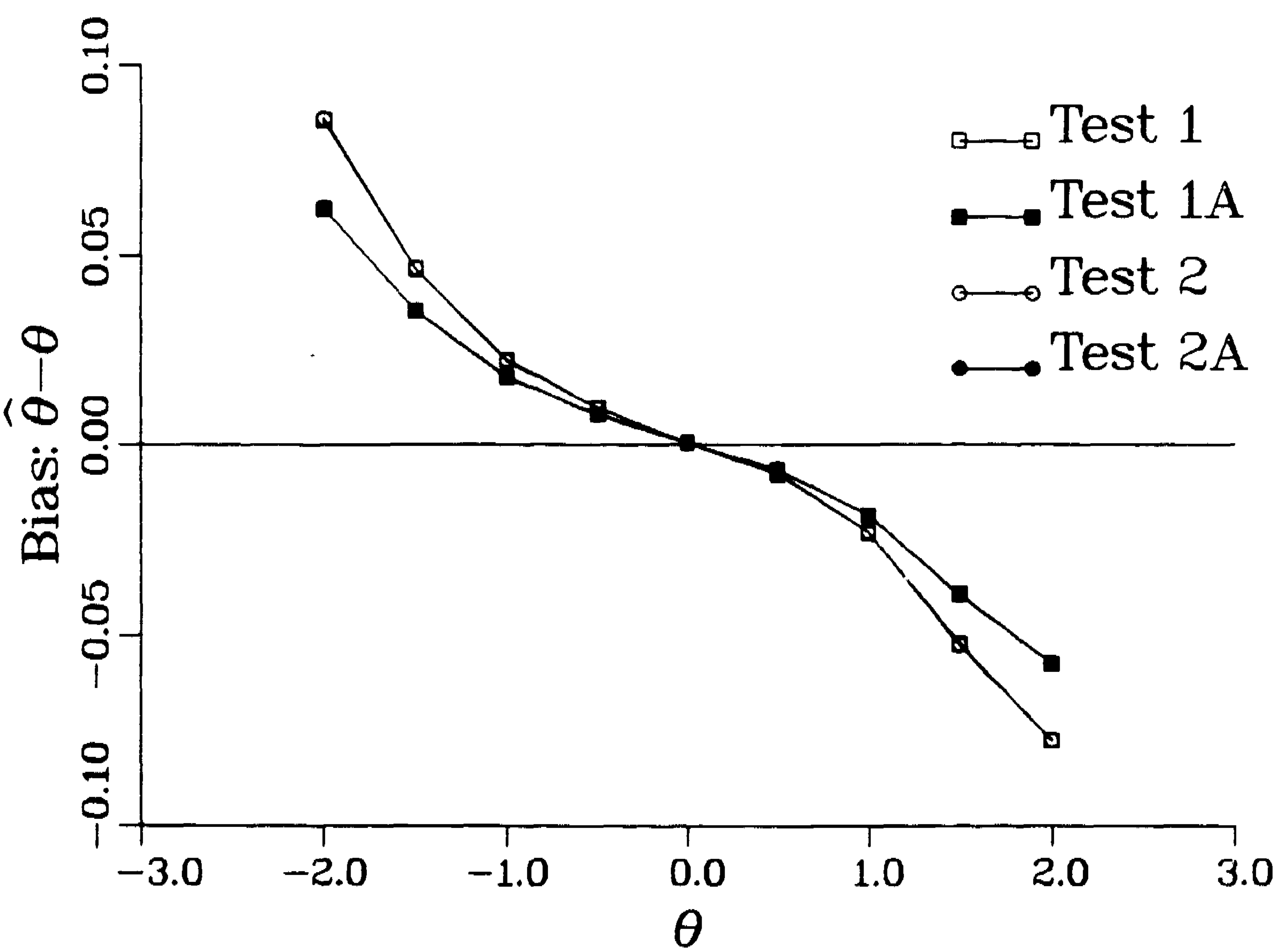
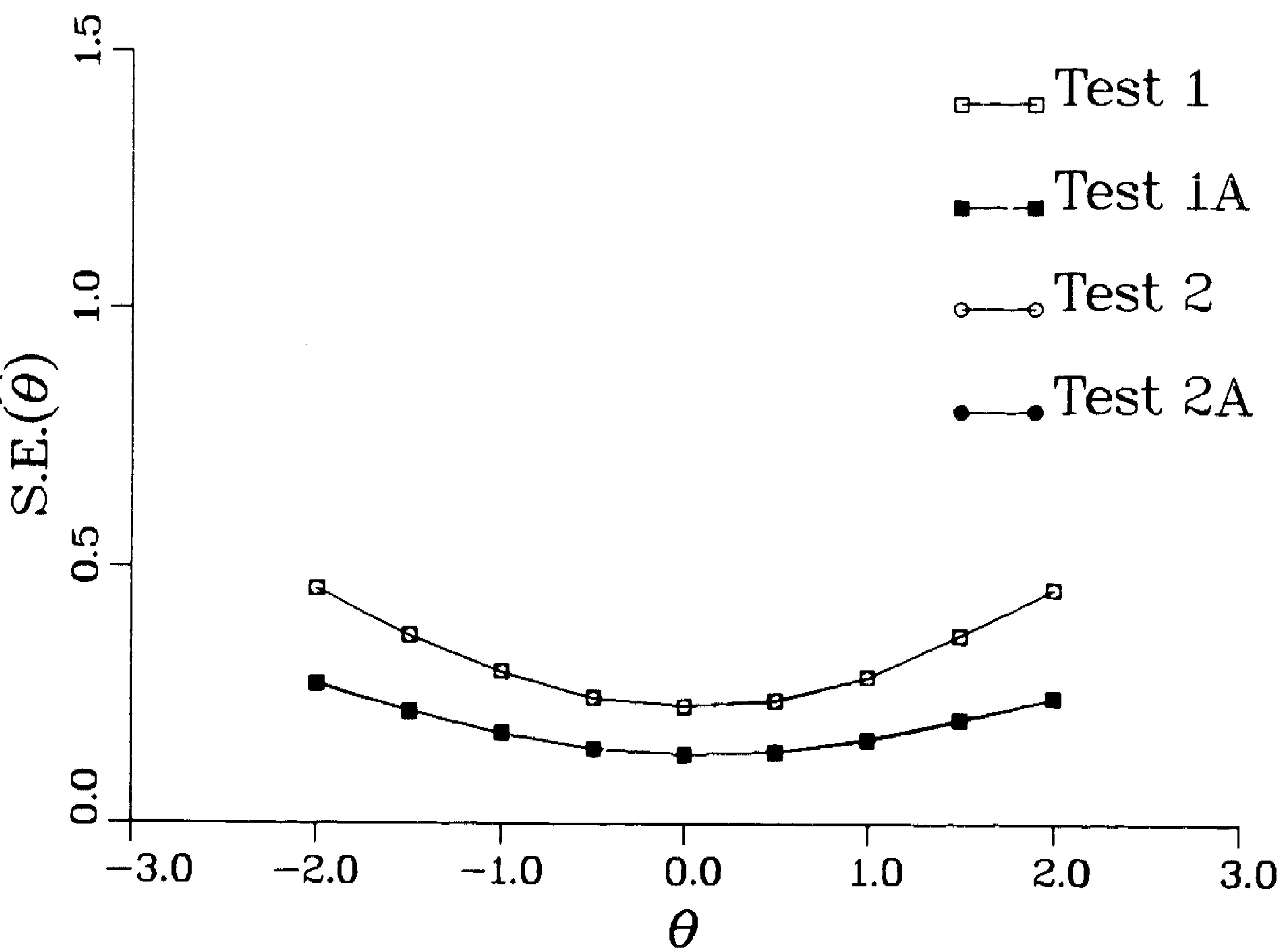

Test 1
Test 1A
Test 2
Test 2A
0.10
0.05
0.00
-0.05
-0.10
Bias: $\hat{\theta}-\theta$
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
θ

Figure 4



Test 1
Test 1A
Test 2
Test 2A
S.E.($\hat{\theta}$)
0.0
0.5
1.0
1.5
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ